Supplementary Information
Improving Polygenic Prediction in Ancestrally Diverse Populations

Yunfeng Ruan[1,2], Yen-Feng Lin[3,4,5], Yen-Chen Anne Feng[1,6,7,8,9,10], Chia-Yen Chen[11], Max Lam[1,8,12,13,14], Zhenglin Guo[1], Stanley Global Asia Initiatives[*], Lin He[2], Akira Sawa[15], Alicia R. Martin[1,8], Shengying Qin[2,60] ✉, Hailiang Huang[1,8,16,60] ✉, Tian Ge[1,6,7,17,60] ✉

[1]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, Shanghai, China. [3]Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan. [4]Department of Public Health & Medical Humanities, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan. [5]Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan. [6]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [7]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [8]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [9]Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan. [10]Master of Public Health Program, National Taiwan University, Taipei, Taiwan. [11]Biogen, Cambridge, MA, USA. [12]Division of Psychiatry Research, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY, USA. [13]Research Division, Institute of Mental Health Singapore, Singapore, Singapore. [14]Human Genetics, Genome Institute of Singapore, Singapore, Singapore. [15]Departments of Psychiatry, Neuroscience, Biomedical Engineering, Genetic Medicine, and Mental Health, Johns Hopkins University School of Medicine and Bloomberg School of Public Health, Baltimore, MD, USA. [16]Department of Medicine, Harvard Medical School, Boston, MA, USA. [17]Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [60]These authors jointly supervised this work: Shengying Qin, Hailiang Huang, Tian Ge. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: chinsir@sjtu.edu.cn (S.Q.); hhuang@broadinstitute.org (H.H.); tge1@mgh.harvard.edu (T.G.)

SUPPLEMENTARY METHODS

The model. PRS-CSx employs the following Bayesian high-dimensional linear regression model for K populations:

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \quad \boldsymbol{\epsilon}_k \sim \mathrm{MVN}\left(\boldsymbol{0}, \sigma_k^2\boldsymbol{I}\right), \quad \pi\left(\sigma_k^2\right) \propto \sigma_k^{-2}, \quad k = 1, 2, \cdots, K,$$

where, for each population k, $\boldsymbol{y}_k$ is a vector of standardized phenotypes (zero mean and unit variance) from N_k individuals, $\boldsymbol{X}_k$ is an $N_k \times M_k$ matrix of standardized genotypes (each column has zero mean and unit variance), $\boldsymbol{\beta}_k$ is a vector of SNP effect sizes, $\boldsymbol{\epsilon}_k$ is a vector of normally distributed non-genetic effects with variance σ_k^2, for which we assign a non-informative prior, and $\boldsymbol{I}$ is an identify matrix. We use $j = 1, 2, \cdots, M$ to index the M unique SNPs across populations. For SNP j in population k, a continuous shrinkage prior is placed on its effect size β_{jk}, which can be represented as global-local scale mixtures of normals:

$$\beta_{jk} \sim \mathrm{N}\left(0, \frac{\sigma_k^2}{N_k}\psi_j\right), \quad \psi_j \sim \mathrm{G}\left(a, \delta_j\right), \quad \delta_j \sim \mathrm{G}(b, \phi),$$

where ϕ is a global shrinkage parameter shared across all SNPs that models the overall sparseness of the genetic architecture, and ψ_j is a local, SNP-specific shrinkage parameter that is adaptive to marginal GWAS associations. Since both ϕ and ψ_j do not depend on k, the continuous shrinkage prior is shared across populations. Note that for any $c > 0$, if a random variable X follows a gamma distribution, $X \sim \mathrm{G}(\zeta, \eta)$, then $cX \sim \mathrm{G}(\zeta, \eta/c)$. The variance of β_{jk} thus scales with both ϕ and ψ_j.

Full conditionals. Let $\mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $\mathrm{G}(\zeta, \eta)$ and $\mathrm{iG}(\zeta, \eta)$ denote the gamma distribution and inverse gamma distribution, respectively, with probability density functions

$$f_{\mathrm{G}}(x; \zeta, \eta) = \frac{\eta^\zeta}{\Gamma(\zeta)}x^{\zeta-1}\exp(-\eta x), \quad f_{\mathrm{iG}}(x; \zeta, \eta) = \frac{\eta^\zeta}{\Gamma(\zeta)}x^{-\zeta-1}\exp\left(-\frac{\eta}{x}\right), \quad x > 0, \quad \zeta > 0, \quad \eta > 0,$$

where $\Gamma(\cdot)$ is the gamma function. Let $\mathrm{giG}(\lambda, \rho, \chi)$ denote the three-parameter generalized inverse Gaussian distribution with density function

$$f_{\mathrm{giG}}(x; \lambda, \rho, \chi) = \frac{(\rho/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\rho\chi})}x^{\lambda-1}\exp\left\{-\frac{1}{2}\left(\rho x + \frac{\chi}{x}\right)\right\}, \quad x > 0, \quad \rho > 0, \quad \chi > 0,$$

where K_λ is the modified Bessel function of the second kind. In addition, let $\widehat{\boldsymbol{\beta}}_k = \boldsymbol{X}_k^T\boldsymbol{y}_k/N_k$ denote the marginal least squares effect size estimates for population k and $\boldsymbol{D}_k = \boldsymbol{X}_k^T\boldsymbol{X}_k/N_k$ denote the LD matrix for population k. $\boldsymbol{\Psi} = \mathrm{diag}\{\psi_1, \psi_2, \cdots, \psi_M\}$ is a diagonal matrix, and K_j is the number of populations in which SNP j is present. The full conditional distributions for unknown model parameters are analytically tractable as shown below.

Posterior distribution of the unknown model parameters:

$$\begin{aligned}\pi\left(\boldsymbol{\beta}_k, \sigma_k^2, \psi_j, \delta_j \mid \boldsymbol{y}_k, k = 1, 2, \cdots, K\right) &\propto \prod\nolimits_{k=1}^{K}\left[f\left(\boldsymbol{y}_k \mid \boldsymbol{\beta}_k, \sigma_k^2\right)\pi\left(\boldsymbol{\beta}_k \mid \sigma_k^2, \psi_j\right)\pi\left(\sigma_k^2\right)\right]\pi\left(\psi_j \mid \delta_j\right)\pi(\delta_j) \\ &\propto \prod\nolimits_{k=1}^{K}\left[\mathrm{MVN}\left(\boldsymbol{y}_k; \boldsymbol{X}_k\boldsymbol{\beta}_k, \sigma_k^2\boldsymbol{I}\right)\mathrm{MVN}\left(\boldsymbol{\beta}_k; \boldsymbol{0}, \frac{\sigma_k^2}{N_k}\boldsymbol{\Psi}\right)\sigma_k^{-2}\right]\mathrm{G}\left(\psi_j; a, \delta_j\right)\mathrm{G}\left(\delta_j; b, \phi\right).\end{aligned}$$

The full conditional distribution of $\boldsymbol{\beta}_k$:

$$\begin{aligned}
\log\left[\pi\left(\boldsymbol{\beta}_k \mid \boldsymbol{y}_k, \sigma_k^2, \psi_j\right)\right] &\propto \log\left[f\left(\boldsymbol{y}_k \mid \boldsymbol{\beta}_k, \sigma_k^2\right)\pi\left(\boldsymbol{\beta}_k \mid \sigma_k^2, \psi_j\right)\right] \\
&\propto \log\left[\mathrm{MVN}\left(\boldsymbol{y}_k; \boldsymbol{X}_k\boldsymbol{\beta}_k, \sigma_k^2\boldsymbol{I}\right)\mathrm{MVN}\left(\boldsymbol{\beta}_k; \boldsymbol{0}, \frac{\sigma_k^2}{N_k}\boldsymbol{\Psi}\right)\right] \\
&\propto -\frac{1}{2\sigma_k^2}\left(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_k\right)^T\left(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_k\right) - \frac{N_k}{2\sigma_k^2}\boldsymbol{\beta}_k^T\,\boldsymbol{\Psi}^{-1}\boldsymbol{\beta}_k \\
&\propto -\frac{N_k}{2\sigma_k^2}\boldsymbol{\beta}_k^T\left(\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{N_k} + \boldsymbol{\Psi}^{-1}\right)\boldsymbol{\beta}_k + \frac{1}{\sigma_k^2}\boldsymbol{y}_k^T\boldsymbol{X}_k\boldsymbol{\beta}_k \\
&\propto \log[\mathrm{MVN}(\boldsymbol{\beta}_k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]
\end{aligned}$$

where

$$\boldsymbol{\Sigma}_k = \frac{\sigma_k^2}{N_k}\left(\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{N_k} + \boldsymbol{\Psi}^{-1}\right)^{-1} = \frac{\sigma_k^2}{N_k}(\boldsymbol{D}_k + \boldsymbol{\Psi}^{-1})^{-1},$$

$$\boldsymbol{\mu}_k = \frac{1}{\sigma_k^2}\boldsymbol{\Sigma}_k\boldsymbol{X}_k^T\boldsymbol{y}_k = \frac{1}{N_k}(\boldsymbol{D}_k + \boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{X}_k^T\boldsymbol{y}_k = (\boldsymbol{D}_k + \boldsymbol{\Psi}^{-1})^{-1}\widehat{\boldsymbol{\beta}}_k.$$

The full conditional distribution of σ_k^2:

$$\begin{aligned}
\pi\left(\sigma_k^2 \mid \boldsymbol{y}_k, \boldsymbol{\beta}_k, \psi_j\right) &\propto f\left(\boldsymbol{y}_k \mid \boldsymbol{\beta}_k, \sigma_k^2\right)\pi\left(\boldsymbol{\beta}_k \mid \sigma_k^2, \psi_j\right)\pi\left(\sigma_k^2\right) \\
&\propto \mathrm{MVN}\left(\boldsymbol{y}_k; \boldsymbol{X}_k\boldsymbol{\beta}_k, \sigma_k^2\boldsymbol{I}\right)\mathrm{MVN}\left(\boldsymbol{\beta}_k; \boldsymbol{0}, \frac{\sigma_k^2}{N_k}\boldsymbol{\Psi}\right)\sigma_k^{-2} \\
&\propto \sigma_k^{-N_k}\exp\left\{-\frac{1}{2\sigma_k^2}\left(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_k\right)^T\left(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_k\right)\right\}\cdot\sigma_k^{-M_k}\exp\left\{-\frac{N_k}{2\sigma_k^2}\boldsymbol{\beta}_k^T\,\boldsymbol{\Psi}^{-1}\boldsymbol{\beta}_k\right\}\sigma_k^{-2} \\
&\propto \left[\sigma_k^2\right]^{-\frac{1}{2}(N_k+M_k)-1}\exp\left\{-\frac{N_k}{2\sigma_k^2}\left[1 - 2\boldsymbol{\beta}_k^T\widehat{\boldsymbol{\beta}}_k + \boldsymbol{\beta}_k^T(\boldsymbol{D}_k + \boldsymbol{\Psi}^{-1})\boldsymbol{\beta}_k\right]\right\} \\
&\propto \mathrm{IG}\left(\sigma_k^2; \frac{N_k + M_k}{2}, \frac{N_k}{2}\left[1 - 2\boldsymbol{\beta}_k^T\widehat{\boldsymbol{\beta}}_k + \boldsymbol{\beta}_k^T(\boldsymbol{D}_k + \boldsymbol{\Psi}^{-1})\boldsymbol{\beta}_k\right]\right).
\end{aligned}$$

The full conditional distribution of ψ_j:

$$\begin{aligned}
\pi\left(\psi_j \mid \boldsymbol{\beta}_k, \sigma_k^2, \delta_j\right) &\propto \prod\nolimits_{k=1}^{K}\left[\pi\left(\beta_{jk} \mid \sigma_k^2, \psi_j\right)\right]\pi\left(\psi_j \mid \delta_j\right) \\
&\propto \prod\nolimits_{k=1}^{K}\mathrm{N}\left(\beta_{jk}; 0, \frac{\sigma_k^2}{N_k}\psi_j\right)\mathrm{G}\left(\psi_j; a, \delta_j\right) \\
&\propto \prod\nolimits_{k=1}^{K}\psi_j^{-\frac{1}{2}}\exp\left\{-\frac{N_k}{2\sigma_k^2}\beta_{jk}^2\psi_j^{-1}\right\}\psi_j^{a-1}\exp\left\{-\delta_j\psi_j\right\} \\
&\propto \psi_j^{a-\frac{K_j}{2}-1}\exp\left\{-\psi_j^{-1}\sum\nolimits_{k=1}^{K}\frac{N_k}{2\sigma_k^2}\beta_{jk}^2 - \delta_j\psi_j\right\} \\
&\propto \mathrm{giG}\left(\psi_j; a - \frac{K_j}{2}, 2\delta_j, \sum\nolimits_{k=1}^{K}\frac{N_k}{\sigma_k^2}\beta_{jk}^2\right).
\end{aligned}$$

The full conditional distribution of δ_j:

$$\begin{aligned}\pi(\delta_j \mid \psi_j) &\propto \pi(\psi_j \mid \delta_j)\pi(\delta_j)\\ &\propto \mathrm{G}(\psi_j; a, \delta_j)\mathrm{G}(\delta_j; b, \phi)\\ &\propto \delta_j^a \exp\{-\psi_j\delta_j\}\delta_j^{b-1}\exp\{-\phi\delta_j\}\\ &\propto \delta_j^{a+b-1}\exp\{-\delta_j(\psi_j+\phi)\}\\ &\propto \mathrm{G}(\delta_j; a+b, \psi_j+\phi).\end{aligned}$$

Gibbs sampler. In summary, the Gibbs sampler for the PRS-CSx model involves the following steps in each Markov Chain Monte Carlo (MCMC) iteration:

- Update $\boldsymbol{\beta}_k$ for each population k:

$$[\,\boldsymbol{\beta}_k \mid \sigma_k^2, \boldsymbol{\Psi}, \widehat{\boldsymbol{\beta}}_k, \boldsymbol{D}_k\,] \sim \mathrm{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \boldsymbol{\mu}_k = \frac{N_k}{\sigma_k^2}\boldsymbol{\Sigma}_k\widehat{\boldsymbol{\beta}}_k, \quad \boldsymbol{\Sigma}_k = \frac{\sigma_k^2}{N_k}(\boldsymbol{D}_k + \boldsymbol{\Psi}^{-1})^{-1},$$

- Update σ_k^2 for each population k:

$$[\,\sigma_k^2 \mid \boldsymbol{\beta}_k, \boldsymbol{\Psi}, \widehat{\boldsymbol{\beta}}_k, \boldsymbol{D}_k\,] \sim \mathrm{iG}\left(\frac{N_k+M_k}{2}, \frac{N_k}{2}\left[1 - 2\boldsymbol{\beta}_k^T\widehat{\boldsymbol{\beta}}_k + \boldsymbol{\beta}_k^T(\boldsymbol{D}_k + \boldsymbol{\Psi}^{-1})\boldsymbol{\beta}_k\right]\right),$$

- Update ψ_j for each variant j:

$$[\,\psi_j \mid \beta_{jk}, \sigma_k^2, \delta_j\,] \sim \mathrm{giG}\left(a - \frac{K_j}{2}, 2\delta_j, \sum\nolimits_{k=1}^{K}\frac{N_k}{\sigma_k^2}\beta_{jk}^2\right),$$

- Update δ_j for each variant j:

$$[\,\delta_j \mid \psi_j\,] \sim \mathrm{G}(a+b, \psi_j+\phi).$$

In practice, for each population k, the genome is divided into independent LD blocks. Posterior effect sizes within each block is updated sequentially within each MCMC iteration.

SIMULATIONS

In the primary simulation, PRS-CS was more accurate than LDpred2 in both within- and cross-population prediction when the discovery GWAS was well-powered, while LDpred2 was more accurate when the discovery sample size was limited, which likely reflects the strengths and limitations of the continuous shrinkage prior vs. the spike-and-slab prior used in PRS-CS and LDpred2, respectively. Specifically, to ensure that posterior effect size estimates are not inflated, we have imposed a minimal shrinkage to the marginal effect size estimates in PRS-CS, which may induce a stronger-than-optimal regularization when the GWAS sample size is small. In contrast, LDpred2 does not guarantee that the bounds of the posterior effect size estimates are finite, which may better separate signals from noise when the GWAS has limited power. However, this comes at the expense of the algorithm being sensitive to imperfectly matched LD reference panels and having convergence issues when the GWAS sample size is large.

We conducted a series of secondary simulations, by varying one parameter in the primary simulation at a time, to assess the generalizability of the observations in the primary simulation and the robustness of PRS-CSx across a wide range of genetic architectures, cross-population genetic overlaps and discovery GWAS sample

sizes. (i) We varied the polygenicity of the genetic architecture by randomly sampling 0.1% or 10% of the HapMap3 variants as causal variants. For fixed heritability, the predictive performance of all PRS construction methods reduced as the genetic architecture became more polygenic, due to the increasing difficulty of accurately estimating small genetic effects and separating signals from noise. The coupled shrinkage prior provided larger gain in prediction accuracy when the genetic architecture was sparse but its benefit was reduced in the extreme polygenic case (Extended Data Fig. 1; Supplementary Table 2). (ii) We varied the cross-population genetic correlation using r_g=0.4 or r_g=1.0. As expected, cross-population prediction accuracy was higher when SNP effect sizes were highly concordant across populations and decreased when genetic effects became less correlated, making effect size estimates less transferable. However, the improvement of PRS-CSx relative to PRS-CS-mult was largely consistent across different genetic correlations (Extended Data Fig. 2; Supplementary Table 3). (iii) We varied the sample size of the discovery GWAS with the ratio of the EUR vs. non-EUR GWAS sample sizes kept unchanged (50K EUR + 10K non-EUR; 200K EUR + 40K non-EUR; 300K EUR + 60K non-EUR). PRS-CS and PRS-CS-mult were less accurate than LDpred2 and LDpred2-mult when the discovery GWAS were small but outperformed LDpred2-based methods when the GWAS became more powerful. The improvement of PRS-CSx over PRS-CS-mult was robust regardless of the variation in the discovery sample size (Extended Data Fig. 3; Supplementary Table 4). (iv) We varied the ratio of the EUR vs. non-EUR GWAS sample sizes with the total sample size kept constant (120K EUR + 0K non-EUR; 80K EUR + 40K non-EUR; 60K EUR + 60K non-EUR). The prediction in non-EUR populations benefitted substantially from increasing the proportion of non-EUR training samples, and the coupled shrinkage prior provided consistent gain in prediction accuracy as the power of the non-EUR GWAS varied (Extended Data Fig. 4; Supplementary Table 5). (v) We varied the SNP heritability of the simulated trait in different populations (h^2=0.5 and 0.25 in EUR and non-EUR populations respectively, and vice versa). SNP heritability determined the overall predictability of a trait, but the relative performance across different polygenic prediction methods was consistent with the primary simulation (Extended Data Fig. 5; Supplementary Table 6). (vi) We reduced the proportion of causal variants that were shared across populations to 70% or 40% to assess the robustness of PRS-CSx when the modeling assumption was violated. The transferability of PRS decreased due to reduced similarity of the genetic architecture across populations, but PRS-CSx continued to outperform alternative methods, and PRS-CS-mult in particular, suggesting that the method is robust to model misspecification (Extended Data Fig. 6; Supplementary Table 7). (vii) We simulated allele frequency and LD dependent genetic architecture, where variants with lower MAF and variants located in lower LD regions tended to have larger effects on the trait. The predictive performance of different PRS construction methods was highly consistent with the primary simulation, suggesting that the methods examined are not sensitive to the coupling between effect size, MAF and LD (Extended Data Fig. 7; Supplementary Table 8). (ix) Lastly, we evaluated the impact of two hyper-parameters, which determined the shape of the continuous shrinkage prior, on the predictive performance of PRS-CSx. We confirmed that the default values of the two parameters used throughout this work, which were consistent with the default setting of PRS-CS, produced optimal prediction accuracy among a grid of values we assessed (Supplementary Table 9). In summary, while the benefits of using a coupled continuous shrinkage prior varied with simulation designs and may be small in certain scenarios, we concluded that PRS-CSx improved cross-population prediction accuracy relative to alternative methods across a vast majority of the simulation settings and was robust to model misspecification.

MCMC DIAGNOSIS

Convergence of the MCMC samplers employed by Bayesian polygenic prediction methods is often overlooked in the literature, likely because (i) the focus of polygenic prediction is to aggregate genetic effects across the genome into a single score, rather than making inference of the genetic effects of individual variants; and (ii) the sheer size of the model parameters (>1 million for each population) makes traditional model diagnostic methods, which often rely on graphical outputs, difficult to apply. To assess the overall convergence of the Gibbs sampler used in PRS-CSx, for each trait, we selected a few SNPs where we monitored the convergence

of their posterior effect size estimates. Some of the SNPs had strong associations with the trait in multiple populations, while some of the SNPs were null across populations. We ran the PRS-CSx model three times using different random seeds, and assessed the convergence using the Gelman-Rubin convergence diagnostic for multiple chains (Gelman & Rubin; *Stat. Sci.* **7,** 457–472, 1992). All reduction factors across the SNPs we examined were smaller than 1.05, indicating convergence. As an example, Extended Data Fig. 9 shows the trace plots and autocorrelation functions (ACFs) for the posterior effects of rs7412 on low-density lipoprotein cholesterol (LDL-C) when integrating UKBB, BBJ and PAGE GWAS summary statistics using PRS-CSx. This SNP, located within the *APOE* locus on chromosome 19, had extremely strong marginal associations with LDL-C across the three GWAS (all *P*-values <1E-200). Trace plots and ACFs (Extended Data Fig. 9) indicated that the Gibbs sampler achieved reasonable convergence and mixing. Future work is needed to better monitor the behavior of the Markov chain in high-dimensional settings.